Genomics-aided structure prediction

Joanna I. Sułkowska[a,1], Faruck Morcos[a,1,2], Martin Weigt[b], Terence Hwa[a,2], and José N. Onuchic[c,2]

[a]Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093-0374; [b]Laboratoire de Génomique des Microorganismes, UMR 7238, Université Pierre et Marie Curie, 15 rue de l'École de Médecine, 75006 Paris, France; and [c]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005-1827

Contributed by José N. Onuchic, May 9, 2012 (sent for review January 20, 2012)

We introduce a theoretical framework that exploits the ever-increasing genomic sequence information for protein structure prediction. Structure-based models are modified to incorporate constraints by a large number of non-local contacts estimated from direct coupling analysis (DCA) of co-evolving genomic sequences. A simple hybrid method, called DCA-fold, integrating DCA contacts with an accurate knowledge of local information (e.g., the local secondary structure) is sufficient to fold proteins in the range of 1–3 Å resolution.

protein folding | residue contact prediction | contact map estimation | residue-residue coevolution | statistical potentials

Proteins are heteropolymers of amino acids that adopt specific 3D structures to perform designated biological tasks. Enormous experimental efforts have been invested to determine a large number of protein structures. Currently, computational structure prediction methods are reasonably successful in describing interactions among residues close (local) in sequence. Given the limited information for residues that are distant in sequence, success in large-scale structure prediction has depended crucially on known structural motifs available in protein databases. In cases where similarity to proteins of known structures exists, methods like fold recognition and homology modeling (1–3) have been shown as successful and effective, according to the Critical Assessment of Techniques for Protein Structure Prediction (4). Nevertheless, the accuracy of these methods is still in many cases far from the resolution needed to explore protein functions.

Here we introduce a new computational approach that exploits information from the rapidly growing genomic sequences to complement the currently limited structural databases. Over the years, a variety of methods has been used to study co-evolution in protein sequences and estimation of residue contacts with mixed success (5–11). Recently, methods based on direct coupling analysis (DCA) (12) were shown to predict 50–300 non-local contacts to 70–80% accuracy for a variety of protein domains (13). DCA is based purely on protein sequence information. It uses covariance in homologous protein sequences as an input and deduces a direct interaction between residues (12). Those with strong direction interactions are shown to be related to structurally conserved residue-residue contacts in the protein fold (12, 13). As the contacts predicted by DCA recapitulate major features of the native contact maps, we developed a simple hybrid method integrating DCA contacts and detailed local information, to fold proteins of up to about 200 amino acids to within 3 Å of the native structures.

Our methodology is guided by the energy landscape theory (14), which asserts that in a minimally frustrated, funnel-like energy landscape, native contacts are on average favorable and dominant over non-favorable, non-native ones. This drives proteins smoothly toward their native states. Folding simulations, using native contacts in structure-based models (SBM), have been successfully applied to fold large and complex proteins, based on the complete knowledge of the native structures (15). Here, we generalize this methodology adopting attractive non-local interactions only for DCA-predicted contacts. The incomplete and imperfect nature of contact information requires substantial modification of SBM, specifically, the use of statistical potentials to describe the detailed forms of the non-local and local residue-residue interactions (16–21). We term this methodology DCA-fold.

Results and Discussion

A summary of the DCA-fold methodology is depicted in Fig. 1. The list of DCA contacts is first generated by sequence analysis (13) of multiple-sequence alignments of the homologous protein family, which contains the protein sequence whose 3D structure shall be predicted. To search for the native protein conformation, we perform simulated annealing using a coarse-grained molecular dynamics model (22) with a single bead per residue, known as the "Cα model." Other than a harmonic interaction potential along the bonds of the protein backbone, only two types of interactions are allowed between two residues at sequence positions i and j: (i) non-local contact interactions (sequence separation $|j-i|>4$), defined by DCA-predicted contacts and described by a contact potential $V_{\text{contact}}(r_{ij})$ which depends on the inter-residue distance r_{ij}; (ii) local interactions ($|j-i|\leq 4$), approximated by a torsional potential $V_{\text{tor}}(\alpha_i,\tau_i)$ which depends on the Cα dihedral angles (α_i,τ_i) at each position i. In SBM where the native distances and angles are used, $V^{\text{NAT}}_{\text{contact}}(r_{ij})$ and $V^{\text{NAT}}_{\text{tor}}(\alpha_i,\tau_i)$ are taken as Gaussian and harmonic functions, respectively, centered about their native values (22). When native information is not available, $V_{\text{contact}}(r_{ij})$ is approximated by a statistical potential for each DCA pair, with its form depending on the nature of the interacting residues a_i and a_j, the sequence distance along the chain $|j-i|$, and the ranking of the DCA contacts. Similarly, a statistical potential is used to describe local interactions (23) $V_{\text{tor}}(\alpha_i,\tau_i)$. A detailed description of these potentials is presented in *Methods* and the *SI Appendix*. Here, we report the effects of these potentials on the accuracy of the predicted structures.

The effectiveness of DCA-fold is evaluated on proteins with the following characteristics: (i) enriched sequence availability (i.e., >1,000 non-redundant homologous sequences) to ensure the necessary statistics for DCA; (ii) known experimental structures, for performance evaluation and access to native structural information; (iii) diversity in fold type (α, β, α/β) and size (52 to 187 residues); and (iv) diversity in domain families. A set of eight proteins was used to develop the appropriate form of statistical potentials, the set of parameters to model the inter-molecular interactions, and the number of DCA-contacts to include. This general model was then used to fold a set of 15 proteins (Table 1, eight training plus seven test), using parameters obtained from the training set. Fig. 2 and *SI Appendix* (*SI Appendix*, Fig. S1) show the DCA-estimated contact maps used to drive DCA-fold

Author contributions: J.I.S., F.M., T.H., and J.N.O. designed research; J.I.S. and F.M. performed research; M.W. contributed new reagents/analytic tools; J.I.S., F.M., M.W., T.H., and J.N.O. analyzed data; and J.I.S., F.M., M.W., T.H., and J.N.O. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]J.I.S. and F.M. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: jonuchic@rice.edu, hwa@ucsd.edu, or amorcosg@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1207864109/-/DCSupplemental.

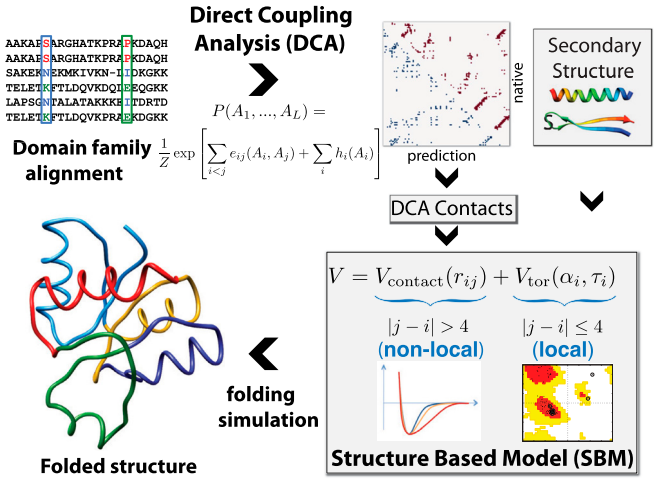

Fig. 1. The DCA-fold methodology: Domain family alignments are used as input for direct coupling analysis (DCA), which generates a large number of accurate contact predictions (13). The DCA contacts are used to drive folding simulations, based on a modified structure-based model (SBM). The Hamiltonian of the SBM contains an interresidue contact potential for local and non-local contacts. Local information is described by a torsional potential together with the local secondary structure, which may be derived from a variety of methods (*SI Appendix* and, for example, ref. 27).

calculations for each protein. The number of DCA contacts used here is based on the optimal number of contacts used for proteins of similar lengths in the training set (Table 1, second column), although the results are insensitive to the choice in this vicinity (*SI Appendix*, Fig. S2).

Our results are expressed in terms of the root-mean-square deviation (Cα-RMSD) of the alignment of predicted positions of Cα atoms with respect to the native ones. The key results, for 80% of the top residues, are summarized in Fig. 3 (with all values shown in Table 1) for various forms of local and non-local potentials used (distinguished by the different symbols/colors). The filled and open symbols refer to the two sets of proteins, training and test set respectively, used in this work (*SI Appendix*, Table S1). We see that while the RMSD generally increases for the larger proteins, the performances obtained for the training and test sets are similar for all cases examined. When the native distances and angles are used (blue squares, with the blue line as a guide), DCA contact maps are able to generate structures with approximately 1 Å RMSDs even for the largest protein. Exemplary structures, shown in Fig. 4, column 1, are indistinguishable from the native ones (*SI Appendix*, Fig. S3). This set of predictions serves to establish the limit of the foldability of the proteins using DCA-generated contact maps. We also analyzed our predictions with the GDT_TS metric (24), which computes the average percentage of residue distances, from the predicted structure with respect to the target structure under different thresholds. Results using this metric are shown in *SI Appendix*, Table S2 and section 7.

With pairwise distances estimated by the statistical potential $V_{\text{contact}}(r_{ij})$ but still retaining the native local information modeled by $V_{\text{tor}}^{\text{NAT}}(\alpha_i, \tau_i)$ (Fig. 3, red circles with the red line to guide the eyes), we obtained RMSDs in the range of 0.7–2.6 Å. Fig. 4, column 2, shows that all the native folds are faithfully captured when compared with the native structures (*SI Appendix*, Fig. S3). These results indicate that DCA-fold is tolerant to approximate distance potentials and suggest that DCA with good local information is sufficient to derive structures to approximately 3 Å resolution, even for the largest protein. As mentioned, a variety of computational methods currently exists to give faithful description of local information; for example, local template modeling (25) and threading (26). Thus, we expect less than 3 Å resolution structures to be achievable by integrating these local methods with DCA-fold. To further evaluate the contribution of the DCA contacts, we ran a series of control simulations using random contact maps (*SI Appendix*, section 9). Each random map has the same number of contacts as the DCA estimates and again using the estimated non-local information $V_{\text{contact}}(r_{ij})$ and the native local information $V_{\text{tor}}^{\text{NAT}}(\alpha_i, \tau_i)$. We observe that most of the predicted structures using random maps had average RMSDs (out of seven random map instances per protein) that are two to three times larger than the ones achieved with the DCA map (*SI Appendix*, Table S3). They are shown with the red × symbols in Fig. 3, with the dashed red line as guide.

If accurate local information is not available, good structures can still be obtained by DCA-fold using the statistical torsional

Table 1. Performance of predicted protein structures with respect to experimental structures

Non-local information			Native	Estimated	Native	Native	Estimated	Estimated
Local information			Native	Native	Estimated (native SS)	Estimated (estimated SS)	Estimated (native SS)	Estimated (estimated SS)
PDB ID (fold)		Length/no. DCA contacts	RMSD in Å for 80% (100%) of residues					
Training set	3nnr (α)	53/114	0.3 (0.4)	0.7 (0.9)	0.5 (1.0)	0.7 (1.2)	1.5 (2.1)	2.1 (2.9)
	1or7 (α)	70/203	0.3 (0.4)	1.0 (1.3)	0.5 (1.0)	1.1 (2.3)	1.9 (2.4)	3.4 (4.4)
	3df8 (α)	91/62	0.6 (0.8)	1.4 (1.8)	2.2 (3.0)	4.2 (5.3)	3.2 (4.2)	5.4 (6.8)
	1oap (α/β)	98/253	0.2 (0.2)	1.0 (1.5)	1.3 (2.3)	1.8 (2.6)	2.5 (3.0)	4.0 (5.2)
	3d7i (α)	98/76	0.9 (1.1)	2.0 (2.7)	1.6 (2.3)	2.6 (3.4)	3.9 (5.2)	4.4 (5.7)
	2gj3 (α/β)	118/220	0.3 (0.4)	1.9 (2.7)	1.3 (2)	1.5 (2.8)	3.7 (4.5)	4.2 (5.5)
	3ddv (β)	139/264	0.4 (0.5)	2.3 (3.0)	1.6 (2.3)	2.8 (4.1)	4.7 (6.1)	5.8 (7.2)
	3nkh (α/β)	187/254	1.0 (1.2)	2.6 (3.3)	2.4 (3.9)	1.9 (4.0)	4.4 (6.2)	5.5 (7.2)
Test set	1jft (α)	54/114	0.3 (0.4)	1.2 (1.4)	1.0 (1.8)	1.3 (1.9)	2.3 (2.8)	3.0 (3.7)
	3f52 (α)	56/11	0.3 (0.4)	1.2 (1.5)	0.7 (1)	0.9 (1.3)	2.6 (3.0)	2.8 (3.3)
	1kgs (α/β)	112/219	0.3 (0.3)	1.8 (2.5)	0.6 (0.9)	1.4 (2.3)	2.0 (2.7)	4.5 (5.9)
	3nyy (β)	112/237	0.4 (0.5)	2.5 (3.0)	1.6 (2.7)	2.2 (3.4)	4.5 (6.0)	5.3 (6.0)
	3fwz (α/β)	116/271	0.3 (0.4)	1.7 (2.3)	0.8 (1.0)	1.6 (2.5)	3.8 (5.1)	4.4 (5.8)
	3fms (α)	120/301	0.4 (0.5)	2.0 (2.6)	0.6 (0.8)	0.7 (1.0)	2.1 (2.9)	2.4 (3.3)
	3bvp (α/β)	132/301	0.4 (0.5)	2.5 (3.6)	1.1 (1.6)	2.6 (3.5)	3.7 (5.0)	5.6 (7.7)

Each column corresponds to different degrees of estimated parameters in the model. The estimated parameters are contact maps based on DCA, the residue pair distance potential (non-local information), and the torsional angle potential (local information). Torsional angle potential has two additional forms: first, when knowledge of secondary structure (SS) is used to guide (α_i, τ_i) estimation; and, second, when a secondary structure prediction tool is used to guide (α_i, τ_i) estimation. The first eight proteins shown in Table 1 were used to refine parameters in the prediction model. The RMSD values in parentheses show the performance of the prediction for 100% of the residues in the predicted structure.

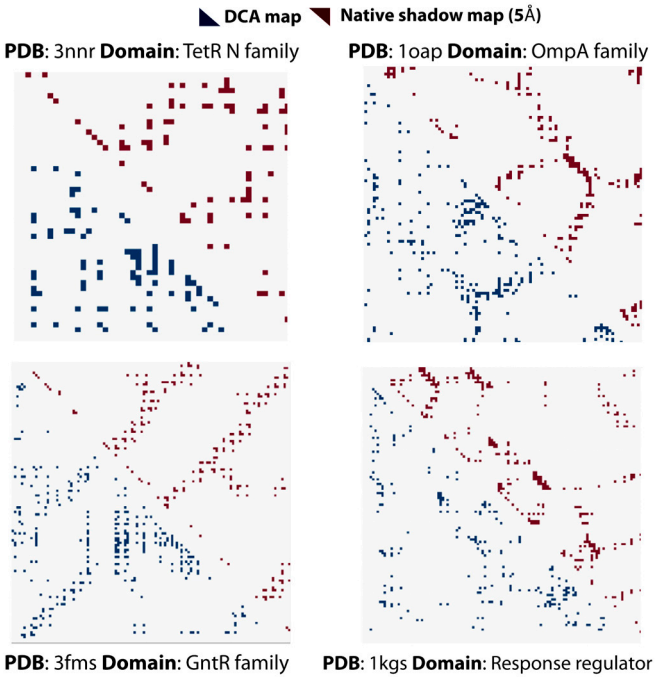

Fig. 2. Comparison of estimated contact maps with native maps for four exemplary proteins (*SI Appendix*, Fig. S1 for maps of all proteins studied). Lower triangular maps, below diagonal, represent DCA contact maps and upper triangular maps are native maps with cutoff value of 5 Å. The prediction results shown in Figs. 3 and 4 and Table 1 used as input a set of contacts estimated using DCA. DCA produces high-quality estimates of contact maps, both in terms of true positive predictions but also in terms of the sparsity of the predicted contacts. Other statistical methods, like mutual information, produce a relatively good number of true positive contacts, but they tend to cluster in specific regions that obscure the global structure of the native contact map (13).

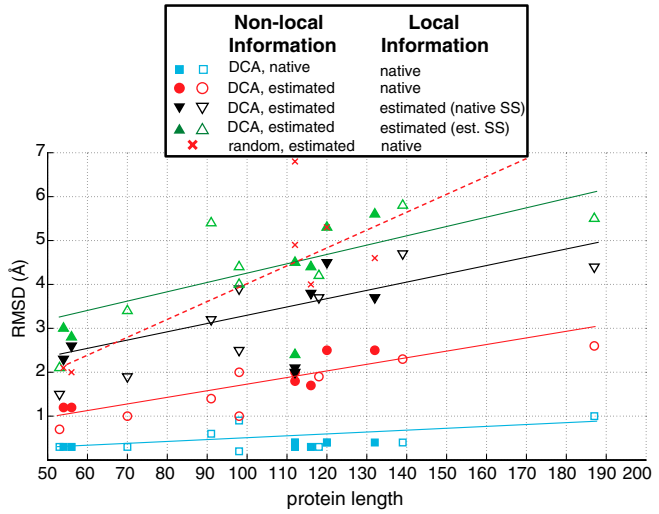

Fig. 3. Predicted RMSDs for 15 proteins of different sizes. The symbols indicate the nature of the information on local and non-local residue interactions. The results shown here correspond to the RMSD for 80% of the residues in the protein in order to avoid the effect of outliers (Table 1 and *SI Appendix*, section 7). Non-local interactions are derived from DCA contacts or random maps (control predictions indicated by symbol ×). Local information is obtained from the native structure or is estimated based on the local secondary structure (SS) classification. The SS classification (α helix or β strand) is obtained via the native structure or is estimated from patterns of the DCA contact map (*SI Appendix*, section 4.3). Open symbols refer to proteins used to derive the statistical potentials, while filled symbols refer to proteins that were used to test this model. The lines are guides to trends by symbols of the same color.

potential $V_{\text{tor}}(\alpha_i, \tau_i)$, together with local secondary structure information. In this scheme, $V_{\text{tor}}(\alpha_i, \tau_i)$ is forced into one of the two known forms, depending on whether the residue i is classified as a part of an α-helix or β-strand (*SI Appendix*, section 4.3). Using a simple secondary structure predictor, which we developed based solely on the incomplete DCA contact maps (*SI Appendix*, Fig. S1), we obtained average RMSD of 4.2 Å in the range of 2.1–5.8 Å for all proteins (Fig. 3, green upper triangles and the green line). Given that secondary structure prediction is a well-developed field (27), we also tested how much can be gained from more accurate secondary structure predictions, by assigning the secondary structure element each residue belongs to based on the native structure (but, not using the native torsional angles; *SI Appendix*, section 4.2). The results (Fig. 3, black lower triangles) improved substantially, with individual proteins gaining as many as 3 Å in RMSD compared to the green triangles. For this case, the average RMSD is 3.1 Å (1.5–4.7 Å). Visual examination of the exemplary structures in Fig. 4 (compare columns 1 and 3 with native structures in *SI Appendix*, Fig. S3) confirms the faithful reconstruction of all the major structure elements. Thus, DCA-fold with approximate local information $V_{\text{tor}}(\alpha_i, \tau_i)$ and an accurate secondary structure predictor can already provide useful protein structure estimates.

All the results discussed so far are with the RMSD computed for the top 80% of the residues. The same general trend is found if we use 100% of the residues (*SI Appendix*, Fig. S4) with values shown in the parenthesis of Table 1. The average increase of approximately 40% in RMSD evidently arises from a small group of 15–20% of residues. One possible explanation is the lack of predicted contacts near the termini of some of the predicted proteins. These effects are analyzed in more detail for two individual proteins in *SI Appendix* (*SI Appendix*, Figs. S5 and S6 and section 7.1.1).

All of the results are stable to atomic scale relaxation (*SI Appendix*, section 6 and Table S4), which affected the RMSD up to 1 Å. Thus, the predicted Cα structures are free of steric clashes, another confirmation of the soundness of the predicted structures. Our results can be further improved by optimizing the parameters of the statistical potentials for individual proteins, leading to as much as 50% gains over the resolutions reported in Table 1, especially for the large proteins (*SI Appendix*, Tables S5 and S6).

Conclusion

The results obtained by DCA-fold demonstrate the power of incorporating genomic sequence information into current structure prediction methods. If all local information on intermolecular interaction is available, then DCA contacts driving simple Cα models can generate structures in the range of 0.7–2.6 Å accuracy (average of 1.7 Å) even for complex proteins. Approximate local information derived from the local secondary structures can already infer structures to the range of 1.5–4.7 Å (average 3.1 Å) for the same proteins. Thus, DCA-fold provides a powerful framework that can significantly enhance the performances of the state-of-the-art structure prediction methods by providing constraints from a large number of reliable non-local contacts derived from genomic sequences via DCA.

This finding is corroborated by an independent analysis (28), published during the submission of this work. In ref. 28, the authors first inferred a contact map using DCA and independently made secondary structure predictions. Subsequently, they predicted 3D protein structures by applying a distance minimization algorithm to embed the DCA-predicted contact maps into 3D, aided by their predicted secondary structures, as well as other heuristics. Marks et al. did not systematically investigate the relative importance of the different kinds of information employed, e.g., the long-range DCA-inferred contacts and the independently inferred local secondary structures. Therefore, it is not clear, based on ref. 28 alone, how much of the final resolution

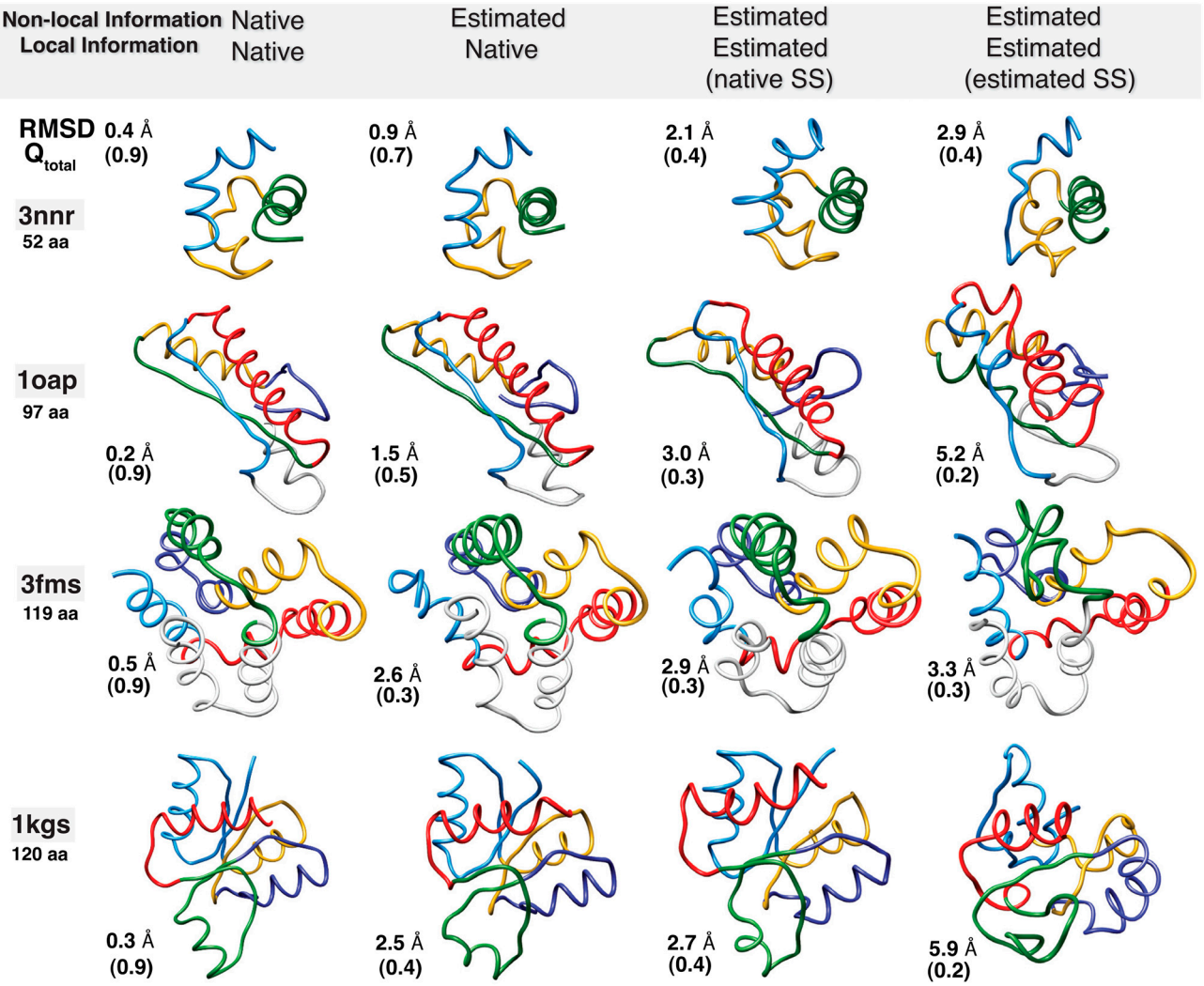

Fig. 4. Protein structures predicted using DCA-fold (Table 1 and *SI Appendix*, Table S1). Their predicted contact maps are shown in Fig. 2. Prediction accuracy for complete proteins is measured in RMSD and by the Q metric (Q_{total}), where the latter characterizes the difference between the predicted and target structures independently of alignment (*SI Appendix*, section 7). These structures are predicted based on DCA contact maps. The results in column 1 were obtained using native contact distances and local information. These structures are indistinguishable from the native ones (*SI Appendix*, Fig. S3). The results in column 2 were obtained using a statistical contact potential and native local information. Columns 3 and 4 show predictions where the local information used was also estimated, based on the type of secondary structure (SS) a residue belongs to. The native SS classification was used in column 3 and a simple SS estimator based on DCA output was used in column 4.

in the final structure obtained resulted from co-evolutionary information. However, since the RMSD values Marks et al. obtained for their sample proteins lie between those we find with local information derived from estimated and native secondary structures (Fig. 3, black and green symbols), we expect their results to be improved toward the solid red line by a more accurate modeling of the local information. The current approach allows such information to be added in a modular manner.

Currently, accurate DCA predictions require approximately 1,000 non-redundant homologous protein sequences (13). This is becoming accessible for many bacterial proteins due to the large number of sequenced bacterial genomes (29), but so far only for few highly amplified eukaryotic proteins. With rapid advances in genomic sequencing capabilities, we however expect DCA-fold may soon be applicable to a wide range of eukaryotic proteins as well.

Methods

Structure-Based Model. We used a structure-based model (SBM) (30), where each amino acid is represented by a single bead of unit mass placed at the location of the Cα atom. The basic form of the potential is $V = V_{\text{contact}}(r_{ij}) + V_{\text{tor}}(\alpha_i, \tau_i)$. The contact potential is composed of three terms:

$$V_{\text{contact}}(r_{ij}) = \sum_{\substack{\text{DCA contacts}\\(i,j>i+4)}} \varepsilon_C \left[\left(1 + \left[\frac{\sigma^C}{r_{ij}}\right]^{12}\right)\left(1 - \exp\left[\frac{-(r_{ij} - r_{ij}^{\text{est}})^2}{2(\sigma_{ij}^{\text{est}})^2}\right]\right) - 1\right] + \sum_{\substack{\text{non contacts}\\(i,j>i+4)}} \varepsilon_R \left[\frac{\sigma^C}{r_{ij}}\right]^{12} + \sum_{\substack{\text{bonds}\\(i,j=i+1)}} k_b (r_{ij} - r_{ij}^b)^2, \quad [1]$$

where the first summation on the right-hand side describes the interaction between non-bonded atoms pairs that are predicted to be in contact based on DCA. The symbol r_{ij}^{est} corresponds to the estimated distance between the pair i, j and width σ_{ij}^{est} of the interacting potential (31). The second term is independent of r_{ij}^{est} and maintains the excluded volume of the polypeptide. The parameter σ^C corresponds to the repulsive size of the beads, between all non-local pairs. The last term represents harmonic interactions between beads adjacent in the sequence, separated by estimated bond distance, r_{ij}^b. The potential V_{tor} describes the local propensity of the chain by a traditional dihedral potential with Cα dihedral angles (α_i, τ_i) at each position i:

$$V_{\mathrm{tor}}(\alpha_i, \tau_i) = \sum_{\mathrm{angle}} k_a(\tau_i - \tau_i^{\mathrm{est}})^2 + \sum_{\mathrm{dihedral}} k_d\Big([1 - \cos(\alpha_i - \alpha_i{}^{\mathrm{est}})] + \frac{1}{2}[1 - \cos(3(\alpha_i - \alpha_i{}^{\mathrm{est}}))]\Big). \quad [2]$$

where (α_i, τ_i) were obtained based on ϕ and Ψ torsional angles defined for N-Cα and C-Cα bonds (*SI Appendix*, section 4.1). Superscript *est* is used to refer to a single estimated parameter as the reference state. The interaction strengths are $k_b = 2 \times 104\ \varepsilon/\mathrm{nm}^2$, $k_a = 40\ \varepsilon/\mathrm{rad}^2$, $k_d = \varepsilon$, $\varepsilon_C = 1\varepsilon$, $\varepsilon_R = 1\varepsilon$ with the reduced unit of energy $\varepsilon_R = k_B T$. This model has been characterized in detail elsewhere (30).

Contact Map Based on DCA Predictions. The prediction of protein structures (eight proteins from the training set and seven test proteins) was performed based on DCA-contact maps and the energy function $V(r_{ij})$, where all parameters (*est*) were obtained from statistical potentials. In order to decide the number of DCA contacts to use as input to DCA-fold, we systematically tested different numbers of DCA contacts for each of the training proteins until we found the optimum prediction. For the testing proteins, we used similar number of DCA contacts as the proteins of similar length in the training set. In general, we observe that the prediction results are robust to the specific number of DCA contacts selected (*SI Appendix*, Fig. S2).

Structure-Based Model Combined with Statistical Potentials. To construct statistical potentials for each DCA pairs we developed a series of distance potentials based on (32–34). The form of these potentials is characterized by a minimum at the estimated pairwise distances based on training set of 65 proteins (35), the frequency of occurrence of different type of interacting residues a_i, a_j, the chemical properties of interacting the amino acids a_i and a_j and their sequence separation $|j - i|$. We optimized the coefficients in our potential in a way to obtain a minimally frustrated landscape based on training set composed of eight proteins. Local interactions $V_{\mathrm{tor}}(\alpha_i, \tau_i)$ were as well-described by statistical potentials. Dihedral angles were estimated based on protein sequences with neighbor dependent probability distributions calculated by Ting el al. (23). These pairwise Ramachandran distributions were combined (23) to get estimates of ϕ and Ψ angles for the all-atom representation. The inferred angle values were in general biased toward alpha helix prediction. This bias was corrected using either the native knowledge of the secondary structure (SS) classification or with a SS prediction algorithm based on DCA contact maps (*SI Appendix*, Algorithm S1). In order for our SS prediction algorithm to determine if a given residue belongs to a secondary structure classification (alpha helix, beta strand, or other), the algorithm searches in the DCA contact map for features that could represent secondary structure elements. If it is known a priori that a given sequence triplet belongs to a β-strand or α-helix, then we restrict the estimates to a preferred quadrant of typical Ramachandran distributions for β-strands or α-helices. This way, the maximization procedure in supplementary Eq. **6** would get the angles with highest probability constrained to such predefined quadrants. For the case of the alpha helix, we constrain the estimates to a region where $\phi < -60$ and $-90 < \Psi < -40$. We defined the corresponding region for beta strands as being $\phi < -100$ and $\Psi > 80$. For the rest of possible configurations, like loops, turns, left-handed alpha helices, etc., we do not constrain the estimation and use the plain formulation in (23).

Predicted values of ϕ and Ψ angles were converted to α_i, τ_i which described conformation of backbone in Cα model based on a relation developed by Levitt (36). Derivation of statistical potentials and a description of the best parameters are presented in detail in *SI Appendix*.

Molecular Dynamics Simulations of Folding. Simulations were performed with the GROMACS 4.0.5 software package (37). Reduced units were used for all calculations with time steps of size 0.0005. We performed stochastic dynamics with annealing protocol and the Nose-Hoover thermostat (38). The annealing protocol was specified as a single sequence of corresponding time steps and reference temperatures. We used a high temperature which strongly favored the unfolding condition and a few very low temperatures, which strongly favored the folded configuration. For each protein, up to 300 annealing runs were performed for each protein, and the best observed snapshot was chosen. As minimization is inherently statistical, this ensures convergence on the resolution of protein structures.

All-Atom Reconstruction and Empirical All-Atom Force Field for Refinement. To reconstruct all heavy atoms in predicted protein conformations, we used PULCHRA software (39). After reconstruction, the predicted structures were additionally relaxed with an empirical all-atom force field for refinement. We used Amber99 (as a force field in GROMACS) with explicit Tip3p solvent and counter ions (40). We used stochastic dynamics with a time step of 2 fs, and Particle Mesh Ewald electrostatics (41).

ACKNOWLEDGMENTS. We thank Hendrik Szurmant, Bryan Lunt, and Peter Wolynes for discussions during the course of this work. This work was supported by the National Science Foundation through the Center for Theoretical Biological Physics (Grant PHY-0822283) and through J.N.O. (NSF-MCB-1214457).

1. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA* 101:7594–7599.
2. Leaver-Fay A, et al. (2011) ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487:545–574.
3. Fiser A, Sali A (2003) Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol* 374:461–491.
4. Moult J, Fidelis K, Kryshtafovych J, Rost B, Tramontano A (2009) CASP8 Proceedings. *Proteins* 77(Suppl 9):1–228.
5. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
6. Fariselli P, Casadio R (1999) A neural network based predictor of residue contacts in proteins. *Protein Eng* 12:15–21.
7. Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng* 14:835–843.
8. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18:309–317.
9. Pollastri G, Baldi P (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 18 (Suppl 1):S62–S70.
10. Shao Y, Bystroff C (2003) Supplement: Fifth Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. *Proteins* 53:497–502.
11. Hamilton N, Burrage K, Ragan MA, Huber T (2004) Protein contact prediction using patterns of correlation. *Proteins* 56:679–684.
12. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72.
13. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108:E1293–E1301.
14. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14:70–75.
15. Kolinski A, Skolnick J (2004) Reduced models of proteins and their applications. *Polymer* 45:511–524.
16. Miyazawa S, Jernigan R (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* 18:534–552.
17. Sun S (1993) Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci* 2:762–785.
18. Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* 15:2507–2524.
19. Tanaka S, Scheraga H (1976) Medium- and long-range interaction parameters between amino acids for predicting 3D structures of proteins. *Macromolecules* 9:945–950.
20. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force—An approach to the knowledge-based prediction of local structures in globular-proteins. *J Mol Biol* 213:859–883.
21. Papoian G, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG (2004) Water in protein structure prediction. *Proc Natl Acad Sci USA* 101:3352–3357.
22. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298:937–953.
23. Ting D, et al. (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput Biol* 6:e1000763.
24. Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31:3370–3374.
25. Wu S, Szilagyi A, Zhang Y (2011) Improving protein structure prediction using multiple sequence-based contact predictions. *Structure* 19:1182–1191.
26. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.
27. Aloy P, Stark A, Hadley C, Russell RB (2003) Predictions without templates: New folds, secondary structure, and contacts in CASP5. *Proteins* 53(Suppl 6):436–456.
28. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6:e28766.
29. Pagani I, et al. (2012) The Genomes OnLine Database (GOLD) v.4: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 40(Database issue):D571–D579.
30. Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: Simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res* 38:W657–W661.

31. Lammert H, Schug A, Onuchic JN (2009) Robustness and generalization of structure-based models for protein folding and function. *Proteins* 77:881–891.
32. Hardin C, Eastwood MP, Prentiss MC, Luthey-Schulten Z, Wolynes PG (2003) Associative memory Hamiltonians for structure prediction without homology: alpha/beta proteins. *Proc Natl Acad Sci USA* 100:1679–1684.
33. Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG (2003) Statistical mechanical refinement of protein structure prediction schemes. II. Mayer cluster expansion approach. *J Chem Phys* 118:8500–8512.
34. Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys J* 85:1145–1164.
35. Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA* 98:10125–10130.
36. Levitt M (1976) A simplified representation of protein conformations rapid simulation of protein folding. *J Mol Biol* 104:59–107.
37. Van der Spoel D, et al. (2005) Gromacs: Fast, flexible, and free. *J Comput Chem* 26:1701–1718.
38. Hoover WG (1985) Canonical dynamics—Equilibrium phase-space distributions. *Phys Rev A* 31:1695–1697.
39. Rotkiewicz P, Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem* 29:1460–1465.
40. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.
41. Essmann U, et al. (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593.